3Д

УДК 004.912

Г.В. Дорохина[1], *В.А. Акчурин*[2]
[1]Институт проблем искусственного интеллекта МОН Украины и НАН Украины,
г. Донецк
[2]Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
sgv@iai.donetsk.ua

Коррекция словарной базы модуля морфологического анализа «РДМА_ИПИИ»

Статья посвящена выявлению и коррекции ошибок словарной базы модуля морфологического анализа РДМА_ИПИИ. В работе сгенерированы правила определения некорректных значений морфологической информации, выполнена классификация видов ошибок и разработаны рекомендации по коррекции словарной базы.

Введение

Обработку естественно-языковых текстов (ЕЯТ) относят к области искусственного интеллекта. Технологии обработки ЕЯТ нашли своё применение в системах машинного перевода, поисковых системах в сети Интернет, роботах-автоответчиках [1] и т.д.

Одним из первых этапов обработки ЕЯТ является морфологический анализ слов, содержащихся в тексте. В настоящее время средства морфологического анализа русскоязычных текстов являются достаточно развитыми – результаты их оценки представлены в [2], [3].

К настоящему времени ИПИИ разработаны модуль декларативного морфологического анализа слов русского языка «РДМА_ИПИИ» [4] и модуль морфологического анализа без словаря [5]. РДМА_ИПИИ в явном виде хранит парадигмы слов – около 3 млн словоформ, синтезированных по словарю А.А. Зализняка [6].

Оценка применения упомянутых модулей в рамках форума «Оценка методов автоматического анализа текста: морфологические парсеры русского языка» показала, что словарная база РДМА_ИПИИ содержит ряд ошибок, часть из которых связана с неверным заданием морфологической информации (МИ) словоформ – набора грамматических характеристик, присущих словоформе. Источником этих ошибок могли служить поэтапное расширение набора грамматических характеристик, используемых в модуле, а также многочисленные процедуры пополнения и корректировки словарной базы, в ходе совершенствования её наполнения.

Наличие ошибок в словарной базе РДМА_ИПИИ влечет за собой некорректные результаты морфологического анализа. Так как словарная база РДМА_ИПИИ является источником для наполнения базы данных модуля морфологического анализа без словаря, ошибки в ней впоследствии распространятся и на результаты бессловарного морфологического анализа.

В связи с этим **актуальной** является разработка методик проверки словарных баз на наличие некорректных МИ и методики корректировки словарной базы.

Объект исследования – словарная база модуля морфологического анализа.

Предмет исследования – корректность морфологической информации.

Цель работы – коррекция словарной базы модуля морфологического анализа РДМА_ИПИИ. Для достижения цели поставлены и решены следующие **задачи**:

– генерация правил выявления некорректных значений МИ на основе теоретических данных и классификация видов ошибок;

– разработка рекомендаций по коррекции словарной базы.

Правила выявления некорректных значений МИ

Для анализа словарной базы на наличие некорректных значений МИ выберем из неё все значения МИ без повторений. В результате количество различных значений МИ составило 1359.

На основе теоретических данных [7], [8] был сформирован набор из 96 правил. Эти правила можно представить в виде двух таблиц. Табл. 1 отражает перечень обязательных и недопустимых грамматических категорий для частей речи. В этой таблице грамматическая категория, обязательная для некоторой части речи, отмечена цифрой «1» на пересечении соответствующего столбца и строки, а недопустимая грамматическая категория для некоторой части речи – цифрой «0».

Пустые ячейки таблицы на пересечении столбца и строки указывают на то, что грамматическая категория не является обязательной для всех словоформ данной части речи, в то же время парадигма слов данной части речи содержит хотя бы одну словоформу, которой присуща указанная грамматическая категория. В табл. 2 представлены правила определения некорректных значений МИ для случаев, соответствующих пустым ячейкам табл. 1.

Таблица 1 – Обязательные и недопустимые категории для частей речи

	Грамматическая категория / Часть речи	1	2	3	4	5	6	7	8	9	10	11	12
		Падеж	Время	Лицо	Степень сравнения	Вид глагола	Тип числительного	Тип местоимения	Залог	Число	Род	Возвратная форма глагола	Одушевленность
1	Наречие	0	0	0		0	0	0	0	0	0	0	0
2	Деепричастие	0	1	0		1	0	0	0	0	0		0
3	Причастие		1			1	0	0	1	1			0
4	Местоимение-прилагательное		0	0	0	0	0		0			0	0
5	Глагол	0			0	1	0	0					0
6	Местоимение-существительное	1	0	0	0	0	0		0			0	0
7	Существительное	1	0	0	0	0	0	0	0			0	
8	Прилагательное		0	0		0	0	0	0			0	
9	Числительное	1	0	0	0	0		0	0			0	

Для последующей ссылки на правила данной таблицы необходимо каждому из них присвоить некоторый идентификатор.

Правилам табл. 1 присвоим двойной номер. Первая часть будет обозначать часть речи, к которой применяется правило, вторая – номер морфологической категории. Таким образом, правило, запрещающее ненулевое значение категории «Падеж» у глаголов будем обозначать П5.1.

Таблица 2 – Правила определения некорректных значений МИ

Часть речи	Условие	Ошибка	№
Прилагательное	Число = Множественное И Род≠0	Определен род во множественном числе прилагательного	1
	Число = Единственное И Род=0	Не определен род в единственном числе прилагательного	2
	Одушевленность=0 И Падеж=В.п. И (Число=Мн. ИЛИ Род=м.р.)	Не определена одушевленность	3
	Одушевленность≠0 И (Падеж≠В.п. ИЛИ Число=Ед. И Род≠м.р.)	Определена одушевленность	4
	Степень сравнения = Сравнительная Степень И Род≠0	Определен род в сравнительной степени прилагательного	5
	Степень сравнения = Сравнительная И Число≠0	Определено число в сравнительной степени прилагательного	6
	Степень сравнения = Сравнительная И Краткость≠0	Определена краткость в сравнительной степени прилагательного	7
	Степень сравнения = Сравнительная И Падеж≠0	Определен падеж в сравнительной степени прилагательного	8
	Краткая форма И Падеж≠0	Определен падеж	9
Числительное	Тип Числительного = Порядковое И Число = Множественное И Род ≠ 0	Определен признак рода	10
	Тип Числительного = Порядковое И Число = Единственное И Род = 0	Не определен признак рода	11
	Тип Числительного = Порядковое И Число = 0	Не определено число	12
	Падеж=В.п. И Одушевленность=0 И (Число=Мн. ИЛИ Род=м.р.)	Не определена одушевленность	13
	Одушевленность≠0 И (Падеж≠В.п. ИЛИ Число=Ед. И Род≠м.р.)	Определена одушевленность	14
Глагол	Вид глагола = Совершенный И Время = Наст. вр.	Настоящее время у глагола совершенного вида	15
	Вид глагола = Несовершенный И Время = Буд.	Будущее время у глагола несовершенного вида	16
	Наклонение = Повелительное И Время≠0	Определено время в повелительном наклонении глагола	17
	Лицо≠0 И Род≠0	Не заданы лицо и род глагола	18
	Лицо = 0 И (Время = Наст. вр. ИЛИ Время = Буд. ИЛИ Наклонение = Повелительное)	Не определено лицо	19
	Время = Прош. вр. И Род = 0	Не определен род	20
	Переходн. = Непереходный И Залог = Страдательный	Неверный залог	21
	Форма глаг. = Возвратная И Залог = Страдательный	Неверный залог	22
Причастие	Залог = 0	Не определен залог	23
	Непереходный И Залог = Страдательный	Неверный залог причастия	24
	Вид=Совершенный И Время≠Прош.вр.	Неверное время	25
	Число=ед. И Род=0	Не определен род	26
	Число=мн. И Род≠0	Определен род	27
	Число=0	Не определено число	28
	НЕ Краткая форма И Падеж=0	Не определен падеж	29
	Краткая форма И Падеж≠0	Определен падеж	30

В табл. 2 знаки равенство нулю («=0») значения некоторой грамматической категории обозначает, что эта категория не определена в анализируемой МИ, а неравенство нулю («≠0») говорит об определенности категории в анализируемой МИ. Ссылки на правила табл. 2 будем делать по их порядковому номеру (4-й столбец). Например, П28.

С применением описанных выше правил (табл. 1, 2) проведена проверка словарной базы РДМА_ИПИИ на наличие некорректных МИ. В результате было выявлено 211 значений МИ и около 44 500 словоформ, требующих корректировки.

Корректировка словарной базы РДМА_ИПИИ

Внесение автоматических изменений в словарную базу может явиться источником новых ошибок. В связи с этим идеология модуля РДМА_ИПИИ требует проверки человеком запланированных изменений.

Так как количество записей, отнесённых к ошибочным, исчисляется десятками тысяч, необходимо автоматизировать процесс классификации некорректных МИ и формирования рекомендаций по корректировке словарной базы. При этом будем использовать следующую методику.

1. Упорядочим таблицу некорректных МИ по убыванию количества словоформ с данной МИ. Назовём её Исходной таблицей МИ. Таблицу словоформ с ошибочными МИ назовём Таблицей словоформ.
2. Скопируем эту таблицу в таблицу, которую назовём Остатком некорректных МИ.
3. Выберем из таблиц 1, 2 правило, согласно которому первый элемент Остатка некорректных МИ является некорректным.
4. Из Исходной таблицы выберем все записи, удовлетворяющие выбранному правилу. Сформулируем рекомендации по коррекции ошибки. Из Таблицы словоформ выберем все записи с данной ошибкой и убедимся, что применение рекомендации по коррекции ошибки устранит ошибку и не приведёт к появлению новых.
5. Добавим правило к Множеству применённых правил.
6. Сформируем Остаток некорректных МИ путём выбора из Исходной таблицы записей, которые не удовлетворяют ни одному из Множества применённых правил.
7. Если Остаток некорректных МИ не пуст и для его первого элемента количество словоформ с данной МИ больше порогового, перейти на шаг 3.

Таблица 3 – Корректировка словарной базы

Правило	Количество словоформ	Рекомендация
П9, П30	38 468	Заменить значение категории падежа на неопределенное.
П1, П10	4 794	Заменить значение категории рода на неопределенное.
П24	899	Набор словоформ разделен по леммам – получено 27 лемм. Из них: – 8 являются исключениями из правила и корректировке не подлежат; – для 1 ошибочно построены страдательные формы причастия (словоформы подлежат удалению); – в 18 ошибочно отнесены к непереходным (заменить значение категории переходности на «переходный»).
П13	123	Заменить значение категории одушевленности на «неодушевленное».
П14	42	Заменить значение категории одушевленности на неопределенное.
П7.11	120	Данная группа содержит словоформы существительных группы pluralia tantum («имеющие только множественное число»). Часть словоформ этой группы определены как имеющие единственное число (подлежат удалению). Для остальных обнулить в МИ признак возвратной формы глагола.
П23	44	Данная группа содержит причастия от леммы «врезать», с которой словарная база работает некорректно.

Результаты применения данной методики отражены в табл. 3, где также указан порядок применения правил и рекомендации по исправлению ошибок. Данные рекомендации позволяют исправить выявленные ошибки.

Выводы

Научная новизна данной работы состоит в следующем.

1. На основе теоретических сведений сгенерированы правила выявления некорректных значений морфологической информации.

2. Разработана методика корректировки словарной базы модуля морфологического анализа.

Практическая значимость работы состоит в применимости сгенерированных правил и методики для выявления, анализа и исправления ошибок в МИ словарных баз систем обработки русскоязычных текстов на морфологическом уровне.

Литература

1. Антонов А. Диалог 2 роботов о всякой ерунде [Электронный ресурс] / Антонов А. – Режим доступа : http://www.roboter.ru/news/arch_spri_08/dialog_08_04_18.htm
2. Оценка методов автоматического анализа текста: морфологические парсеры русского языка [Электронный ресурс] / О. Ляшевская, И. Астафьева, А. Бонч-Осмоловская [и др.] // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегодной Международной конференции «Диалог» (Бекасово, 26 – 30 мая 2010 г.). – Вып. 9 (16). – М. : РГГУ, 2010. – Режим доступа : http://ru-eval.ru/Dialog2010.pdf
3. Форум «Оценка методов автоматического анализа текста: морфологические парсеры русского языка» : Таблицы оценок 2010 [Электронный ресурс]. – Режим доступа : http://ru-eval.ru/tables_index.html
4. Дорохина Г.В. Модуль морфологического анализа слов русского языка / Г.В. Дорохина, А.П. Павлюкова // Искусственный интеллект. – 2004. – № 3. – С. 636-642.
5. Дорохина Г.В. Модуль морфологического анализа без словаря слов русского языка / Г.В. Дорохина, В.Ю. Трунов, Е.В. Шилова // Искусственный интеллект. – 2010. – № 2. – С. 32-36.
6. Зализняк А.А. Грамматический словарь русского языка: словоизменение, около 100 000 слов / А.А. Зализняк. – М. : Русский язык, 1977. – 880 с.
7. Литневская Е. И. Морфология // Русский язык: краткий теоретический курс для школьников [Электронный ресурс] / Е.И. Литневская. – Режим доступа : http://www.gramota.ru/book/litnevskaya.php?part4.htm
8. Розенталь Д.Э. Справочник по правописанию, произношению, литературному редактированию / Розенталь Д.Э., Джанджакова Е.В., Кабанова Н.П. – [2-е изд., дополнен.]. – М. : ЧеРо, 1998. – 400 с.

Г.В. Дорохіна, В.О. Акчурін

Корегування словникової бази модуля морфологічного аналізу «РДМА_ІПШІ»

Статтю присвячено виявленню та корекції помилок словникової бази бібліотеки морфологічного аналізу РДМА_ІПШІ. В роботі сгенеровано правила визначення некоректних значень морфологічної інформації, класифіковано знайдені помилки та розроблено рекомендації щодо корекції словникової бази.

G.V. Dorokhina, V.A. Akchurin

A Morphological Analysis Module «RDMA_IAI» Vocabulary Database Correcting

The article is devoted to vocabulary database of the morphological analysis module «RDMA_IAI» errors detection and correcting. The rules for the incorrect morphological information values search were generated. The founded errors were classified. There were made the recommendations to correct a vocabulary database.

Статья поступила в редакцию 02.07.2010.